Agricultural Systems **29** (1989) 357–370

An Application of the Markov Chain Approach to Forecasting Cotton Yields from Surveys

J. H. Matis,[a] T. Birkett[b] & D. Boudreaux[a]

[a] Department of Statistics, Texas A&M University, College Station, Texas 77843, USA
[b] NASS, USDA, Washington, DC, 20250, USA

(Received 25 February 1988; revised version received 6 July 1988; accepted 6 July 1988)

ABSTRACT

This paper applies a Markov chain approach to forecasting cotton yield from pre-harvest crop data gathered in a large-scale USDA yield survey. Transition matrices for crop condition classes between successive sampling dates were estimated from three years (1981–1983) of baseline data. The estimated average cotton yields for California and for Texas were forecasted for each pre-harvest sampling date in 1984. The forecasting errors were very encouraging for 1984, and a resampling study of the previous years confirms the relatively small forecast error of this procedure. The procedure should be easy to adapt for similar applications, therefore, the Markov chain approach is recommended as a new, useful procedure for crop forecasting from operational survey data.

INTRODUCTION

A Markov chain approach for forecasting crop yield was presented by Matis *et al.* (1985). In this approach, a transition probability matrix was estimated from historical data. The matrix was used to provide forecasted distributions of final crop yield at selected times prior to harvest for various plant and environmental condition classes. Expected yields and associated standard errors were also obtained for the various crop condition classes. The Markov chain approach is nonparametric, thereby requiring less stringent assumptions; moreover, it provides information which is not

Agricultural Systems 0308-521X/89/$03·50

available from standard regression analyses. However, the approach may involve some loss of precision in the forecast, and the potential loss must be evaluated separately for each application. The technique was illustrated by Matis *et al.* (1985) for a data base created from the CERES-Maize computer model, which simulates the growth and development of corn plants.

This paper applies the Markov chain approach to forecasting cotton yield from pre-harvest crop data gathered in the USDA objective yield survey. The present application is different in many ways from the previous application, and it demonstrates the generality and practicality of the basic Markov chain methodology. The paper first describes the data set and then reviews the Markov chain forecasting approach. The results of analyzing cotton data from the objective yield survey using the new methodology are then presented and discussed.

THE DATA BASE AND OBJECTIVES

The USDA gathers yield data on cotton, as well as a number of other crops, in order to predict the yield and also to later estimate the cotton production

TABLE 1
List of Variables Available from the Objective Yield Survey

List	*Symbol*
For 1·83 m Units	
1. Current number of squares	*TOTSQ*
2. Current number of small bolls and blooms	*TOTBM*
For 12·2 m Units	
1. Current number of large unopened bolls	*BOLLUN*
2. Current number of partially opened bolls	*BOLLPT*
3. Cumulative number of burrs and bolls (accumulated over all visits to date)	*BOLLOP*
4. Cumulative number of burrs and bolls on ground (accumulated over all visits to date)	*BOLLGR*
5. Cumulative number of bolls in sample (*BOLLUN* + *BOLLPT* + *BOLLOP* + *BOLLGR*)	*TOTBL*
6. Cumulative weight of harvested bolls (all opened bolls and bolls on the ground are harvested at each visit)	*CUMWT*
7. Cumulative average weight per boll	*WTBOLL*
8. Number of plants	*PLT*
9. Row spacing	*ROWSP*
10. Yield per hectare (constant × *CUMWT*/*ROWSP*)	*Y*

at harvest. This paper analyzes separately cotton objective yield data from two key producing states, California and Texas, over the four year period 1981–1984. Details of the USDA sampling procedures and of the biometrical variables measured in the cotton survey are given in a USDA publication (USDA, 1987). In brief, a separate random sample of fields was selected in each state for each of the four years. Random sampling units were then located within each selected field. For present convenience, these will be aggregated into one large 12·2 m (40 ft) unit consisting of 3·05 m sections from four different rows in the field and one small 1·83 m unit consisting of two 0·915 m sections, each of which is adjacent to a 3·05 m section. Data were gathered from these two aggregated units for each selected field on five

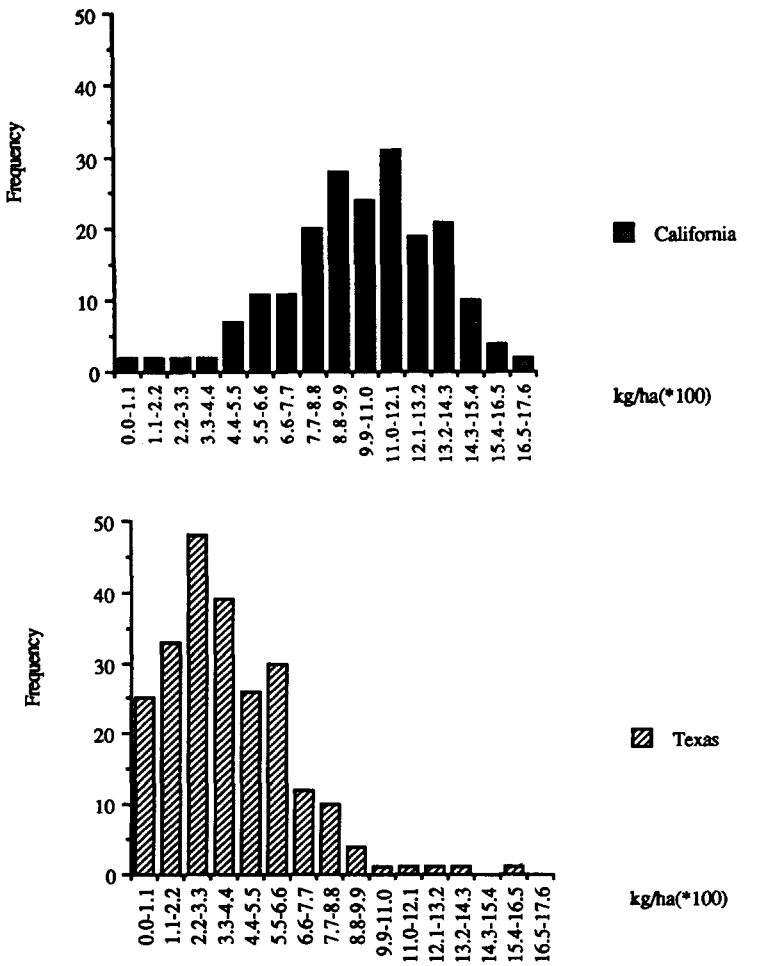

Fig. 1. Histograms of 1984 cotton yields for sampled units in California and Texas.

consecutive months, on dates immediately preceding the first day of August through December.

Table 1 lists the independent variables which were recorded to describe the condition of the cotton plants on each of the five sampling dates. The dependent variable, denoted Y, was the estimated cumulative yield (kg/ha) based on the USDA harvested bolls from the sample unit. The objective of this application was to forecast the estimated average yield of the two states in 1984 from the independent variables available at intermediate times during the 1984 growing season. The forecasting model was developed from a three year (1981–1983) historical data base.

The forecasts were then compared with the final estimates for each state. The distribution of the observed 1984 yields is given in Fig. 1 for each of the two states. There were $n_1 = 196$ fields surveyed in California and $n_2 = 232$ in Texas. The yield estimates, which are to be forecasted, were the means of the observed distributions. These 1984 means were $\bar{Y}_1 = 1036$ kg/ha for California and $\bar{Y}_2 = 403$ kg/ha for Texas.

MARKOV CHAIN PROCEDURE

Simple probability model illustration

The basic theory of the Markov chain procedure was described by Matis *et al.* (1985). The following illustration using a simple probability model may be helpful in clarifying the procedure to establish a framework for the present application. First consider using two of the variables in Table 1; namely, the number of squares (*TOTSQ*) and the number of bolls (*TOTBL*), to describe the condition of a sample unit of cotton plants on 1 August. Discrete plant condition classes may then be defined from these variables. For example, suppose that historically the median values of *TOTSQ* and *TOTBL* on 1 August for cotton plants are 1120 and 83, hence one might define four plant condition classes using the following combinations: (1) $TOTSQ < 1120$ and $TOTBL < 83$, (2) $TOTSQ < 1120$ and $TOTBL \geq 83$, (3) $TOTSQ \geq 1120$ and $TOTBL < 83$, and (4) $TOTSQ \geq 1120$ and $TOTBL \geq 83$. Let us also suppose that for each condition class on 1 August, a probability distribution of cumulative yield at harvest is available. Table 2 lists four such hypothetical distributions. The means of the probability distributions may be used as point predictors of final yield. For example, the predicted yield of cotton in class 1 on 1 August is the weighted average of the class midpoints, i.e. $(0{\cdot}48 \times 550) + (0{\cdot}28 \times 950) + (0{\cdot}17 \times 1200) + (0{\cdot}07 \times 1550) = 842$. The predicted yields of the other classes are 1028, 1106 and 1263, respectively.

TABLE 2
Hypothetical Probability Distributions of Yield for Various Plant Condition Classes in August

Plant condition classes in August		*Yield category (kg/ha)*				*Mean yield (kg/ha)*
		300–800	*800–1 100*	*1 100–1 300*	*1 300–1 800*	
1. $TOTSQ < 1120$	$TOTBL < 83$	0·48	0·28	0·17	0·07	842
2.	$TOTBL \geq 83$	0·26	0·29	0·25	0·20	1 028
3. $TOTSQ \geq 1120$	$TOTBL < 83$	0·19	0·26	0·28	0·27	1 106
4.	$TOTBL \geq 83$	0·08	0·17	0·30	0·45	1 263

This simple example illustrates the basic structure of the Markov chain approach. In practical applications, the probability distributions are estimated from historical data using Markov chain methodology. Some preliminary questions of interest are (1) which variables should be used to define the plant condition classes and (2) how should optimal classes be constructed from these variables. Each of these questions is addressed below with application to the USDA cotton data. In addition to the point estimate, interval estimates, and the size of the forecast error are also relevant and will be addressed.

Selection of variables

The methodology starts with the selection within each given time period of the key independent variables which will be used to define plant condition classes. For some applications, the physiological growth stages serve as a natural series of time periods. However, for the present application, and most other large scale applications, the data are collected on fixed, chronological dates, for example monthly, and the sampling intervals give the time periods. In the present application, data are collected on a few days immediately preceding the first of each month from August to December. For convenience, we will follow the USDA convention of denoting such periods as 'August' through 'December'.

Two regression models were used to assist in selecting the key variables within each period for the baseline (1981–1983) data. One is the multiple linear regression model, the other is the multiple rank regression model in which the independent and dependent variables are transformed into ranks and then analyzed using general linear model theory. The rank regression procedure is nonparametric and focuses on monotonic, as opposed to linear,

TABLE 3
Best Subsets of Two Variables (with R^2 Values) for Regression Models by Month and State

Month	*State*	*Variables in ordinary regression model*	R^2	*Variables in rank regression model*	R^2
August	CA	*TOTSQ, TOTBL*	0·34	*TOTSQ, TOTBM*	0·32
	TX	*TOTSQ, BOLLUN*	0·35	*TOTSQ, TOTBL*	0·31
September	CA	*TOTBL, TOTBM*	0·59	*TOTBL, TOTBM*	0·60
	TX	*BOLLUN, TOTSQ*	0·62	*BOLLUN, TOTSQ*	0·64
October	CA	*CUMWT, BOLLUN*	0·77	*TOTBL, WTBOLL*	0·81
	TX	*CUMWT, BOLLUN*	0·76	*TOTBL, WTBOLL*	0·83

association between variables (Conover, 1980). Best subset regression procedures (SAS, 1982, *PROC RSQUARE*) were utilized for both models. The results are given in Table 3 which lists the best subset of two independent variables for predicting yield, Y, for each state/month combination and for each regression model. In each of these models, a third independent variable was statistically significant, but the increase in R^2 was negligible and of no practical interest. However, the subsequent methodology could easily be generalized to include a third key variable, if desired.

Some subjective judgement was used for the final selection of the key variables from the results in Table 3. *TOTSQ* and *TOTBL* were chosen for

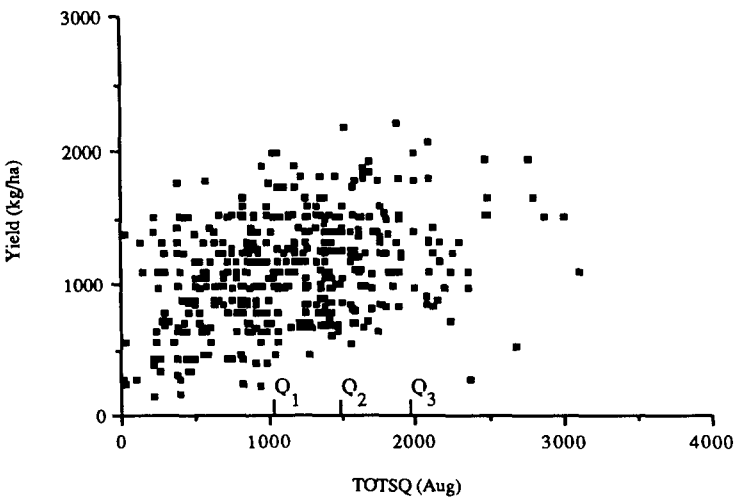

Fig. 2. Scatterplot of yield, Y, verses total square count, *TOTSQ*, in August for 1981–1983 baseline data in California. (A point represents one or more observations and Q_1–Q_3 denote the quartiles).

August for both states since these two variables always had either the highest or next to highest R^2. The highest R^2 for a single variable was 0·24 and for a set of three variables was 0·36, which indicates that two variables are necessary and sufficient for practical purposes in each state. *TOTSQ* and *TOTBL* were also chosen to define plant condition classes in September. Although this combination was not the best in any of the four analyses, it was the second best in each analysis and had the highest overall mean R^2. Moreover, this combination is somewhat robust against early or late growing seasons, *TOTSQ* being an early plant characteristic and *TOTBL* a late one. *TOTBL* and *WTBOLL* were chosen as the optimal set of variables in October for both states on the basis of the preferred, nonparametric regression results.

It should be pointed out that these key variables were selected from the pooled 1981–1983 data. Figure 2 is a scatter plot of pooled data in California for Y vs. *TOTSQ* measured in August. It could be shown that the data also contain a year effect which is small but significant in each regression in each of the three months. This year effect is ignored for the present forecasting, but its effect will be discussed subsequently.

Definition of plant condition classes

The next step in the methodology consists of defining plant condition classes from the key variables. In the simple probability model illustration, each variable was divided in half thereby creating four condition classes. For the present application, the following finer partitioning was implemented for the baseline data. For August, the variable *TOTSQ* for the California data was divided into quarters using the quartiles $Q_1 = 743$, $Q_2 = 1120$, and $Q_3 = 1498$, and *TOTSQ* was divided into halves using the median $M = 83{\cdot}5$; thus there are eight combinations defining plant condition classes. The quartiles for *TOTSQ* are indicated on Fig. 2. For September, the variable *TOTBL* was divided into four parts using the quartiles $Q_1 = 408$, $Q_2 = 573$, and $Q_3 = 573$, and *TOTSQ* was divided using the median $M = 240$, which again gives eight classes. For October, the principal variable, *TOTBL*, was divided into eights due to the higher R^2. The percentiles used for the division were $P_{12{\cdot}5} = 468$, $P_{25} = 571$, $P_{37{\cdot}5} = 655$, $P_{50} = 714$, $P_{62{\cdot}5} = 781$, $P_{75} = 850$ and $P_{87{\cdot}5} = 915$. The other variable, *WTBOLL*, was divided into four classes based on the quartiles $Q_1 = 5{\cdot}80$, $Q_2 = 6{\cdot}37$, and $Q_3 = 7{\cdot}03$, hence for October there are a total of 32 plant condition classes. All of the above specific numbers apply to the California data, and similar classes are defined for the Texas cotton fields. The dependent variable Y was partitioned into 40 discrete classes based on observed quantiles. These classes for the dependent variable are not of equal width, but rather of equal estimated probability.

Transition matrices

Four transition matrices were calculated from the baseline data for each of the states, California and Texas. The first matrix A_{01}, is 1×8 and gives the observed proportions of the 8 condition states defined for August. The second, A_{12}, is an 8×8 matrix. Each row of this matrix sums to 1·0 and contains the observed conditional probabilities that a field in a specified class in August will be in each of the 8 condition classes in September. Similarly, A_{23} and A_{34}, the transition matrices from September to October and from October to final harvest, were 8×32 and 32×40, respectively, and were estimated separately from the 1981–1983 data for California and for Texas.

Predicted yield distributions and yield forecasts for individual fields

The predicted yield distributions may be calculated by multiplying consecutive transition matrices, as proven by Matis *et al.* (1985). The product $A_{01} \cdot A_{12} \cdot A_{23} \cdot A_{34}$ gives a 1×40 matrix which is identical to the observed aggregate yield distribution for the 1981–1983 data. The product $A_{12} \cdot A_{23} \cdot A_{34}$, an 8×40 matrix, gives eight predicted yield distributions, one for each of the eight condition states in August. These eight distributions are analogs of the four hypothetical distributions given in the body of Table 2. Clearly, $A_{23} \cdot A_{34}$ and A_{34} give the predicted yield distributions for September and October, respectively.

The means of these predicted yield distributions may be used as yield forecasts, as illustrated in the example. The means for the previously defined plant condition classes in August, September and October are given in Table 4. The predicted yield distributions contain a wealth of information besides the mean which may be of interest in many applications. For example, other point predictors, such as the median, are easily obtained. Also, forecast intervals, such as 95% prediction intervals, for individual fields may be of great interest in certain applications. Such prediction intervals, which are not constrained by the methodology to be symmetric, are illustrated by Matis *et al.* (1985) but are not of interest in the present application.

RESULTS

Forecasted yield for 1984

Statewide forecasts for 1984 were obtained as follows. A yield forecast was obtained each month for each of the $n_1 = 196$ sample units in California and

TABLE 4
Means (in kg/ha) of Predicted Yield Distributions for Plant Condition Classes by Month

August			*September*		
	Class	*Mean*		*Class*	*Mean*
$TOTBL < M$	$TOTSQ < Q_1$	787	$TOTSQ < M$	$TOTBL < Q_1$	685
	$Q_1 < TOTSQ < Q_2$	915		$Q_1 < TOTBL < Q_2$	926
	$Q_2 < TOTSQ < Q_3$	988		$Q_2 < TOTBL < Q_3$	1135
	$Q_3 < TOTSQ$	1139		$Q_3 < TOTBL$	1359
$TOTBL > M$	$TOTSQ < Q_1$	1109	$TOTSQ > M$	$TOTBL < Q_1$	803
	$Q_1 < TOTSQ < Q_2$	1201		$Q_1 < TOTBL < Q_2$	1085
	$Q_2 < TOTSQ < Q_3$	1275		$Q_2 < TOTBL < Q_3$	1322
	$Q_3 < TOTSQ$	1342		$Q_3 < TOTBL$	1436

October

Class	*Mean*			
	$WTBOLL < Q_1$	$Q_1 < WTBOLL < Q_2$	$Q_2 < WTBOLL < Q_3$	$Q_3 < WTBOLL$
$TOTBL <$ P12·5	437	653	560	693
P12·5 $< TOTBL <$ P25	740	816	866	969
P25 $< TOTBL <$ P37·5	904	990	993	981
P37·5 $< TOTBL <$ P50	948	1042	963	1158
P50 $< TOTBL <$ P62·5	1066	1081	1251	1376
P62·5 $< TOTBL <$ P75	1102	1224	1308	1373
P75 $< TOTBL <$ P87·5	1301	1295	1504	1332
P87·5 $< TOTBL$	1371	1544	1562	1671

$n_2 = 232$ sampled units in Texas by first classifying the individual units into condition classes as defined in the Markov Chain Procedure section and then determining a crop forecast from Table 4. Error-free forecasting would have reproduced the distributions in Fig. 1. The distribution of forecast errors, i.e. actual – forecast, for the individual units in August is given in Fig. 3. In California, the maximum overprediction for the units was 784 kg/ha, and the maximum underprediction was 748 kg/ha. The mean error was 75·3, which corresponds to a statewide forecast of 1111 kg/ha. Since the observed statewide yield was 1036, the simulated forecast error in August for California was 7·2%. The other monthly forecasts and percent errors for California were 1118 (7·9%) in September and 1021 (1·4%) in October. The corresponding forecast results for Texas, with an actual yield of 403 kg/ha, were 385 kg/ha (4·5% error) in August, 398 (1·2%) in September, and 435 (7·9%) in October.

Overall, the results are good by historical standards. The large percentage error in the October forecast in Texas is an outlier which probably indicates

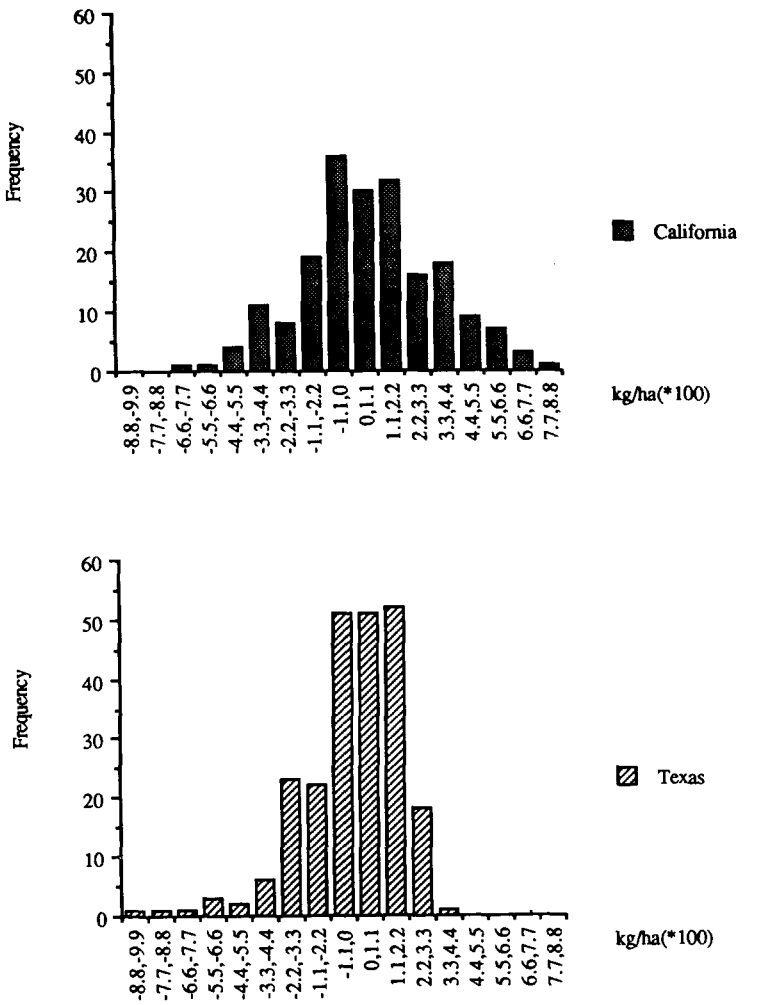

Fig. 3. Histograms of August forecast errors (actual – forecast) of 1984 cotton yields for sampled units in California and Texas.

some weather, insect, or other environmental anomaly affecting yield between October and final harvest.

Investigation of forecast error

In order to further study the forecast error, simulated forecasts were also obtained for 1981, 1982, and 1983. The procedure was repeated for each of these years using the remaining three as baseline data. The actual mean yields for California for the three years were 1150, 1136, and 1020 kg/ha and for Texas were 465, 328, and 352 kg/ha. The key variables were not changed

TABLE 5
Percentage Forecast Errors for Four Separate Years (1981–1984) for each Month/State Combination

	California					*Texas*				
	1981	*1982*	*1983*	*1984*	*Mean*	*1981*	*1982*	*1983*	*1984*	*Mean*
	(*% Forecast error*)					(*% Forecast error*)				
August	3·5	5·8	1·3	7·2	4·5	5·2	0·6	8·1	4·5	4·6
September	3·0	2·1	5·9	7·9	4·7	4·2	0·4	4·4	1·2	2·6
October	1·9	0·2	3·0	1·4	1·6	6·3	3·0	4·4	7·9	5·4
Mean	2·8	2·7	3·4	5·5	3·6	5·2	1·3	5·6	4·6	4·2
(Actual yield)	(1 150)	(1 136)	(1 020)	(1 036)	(1 085)	(465)	(328)	(352)	(403)	(388)

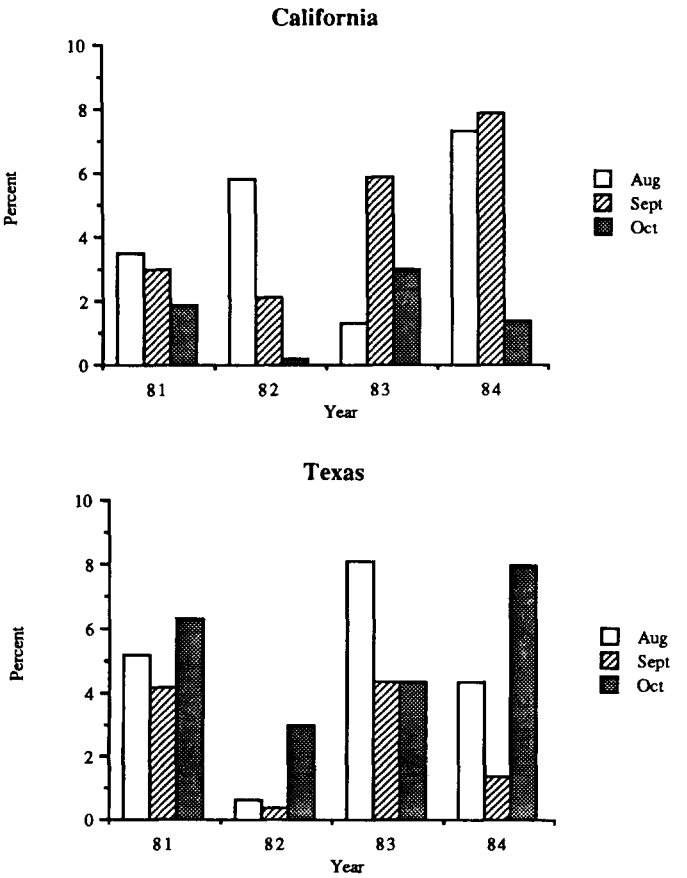

Fig. 4. Comparisons of forecast errors by month within year for California and Texas.

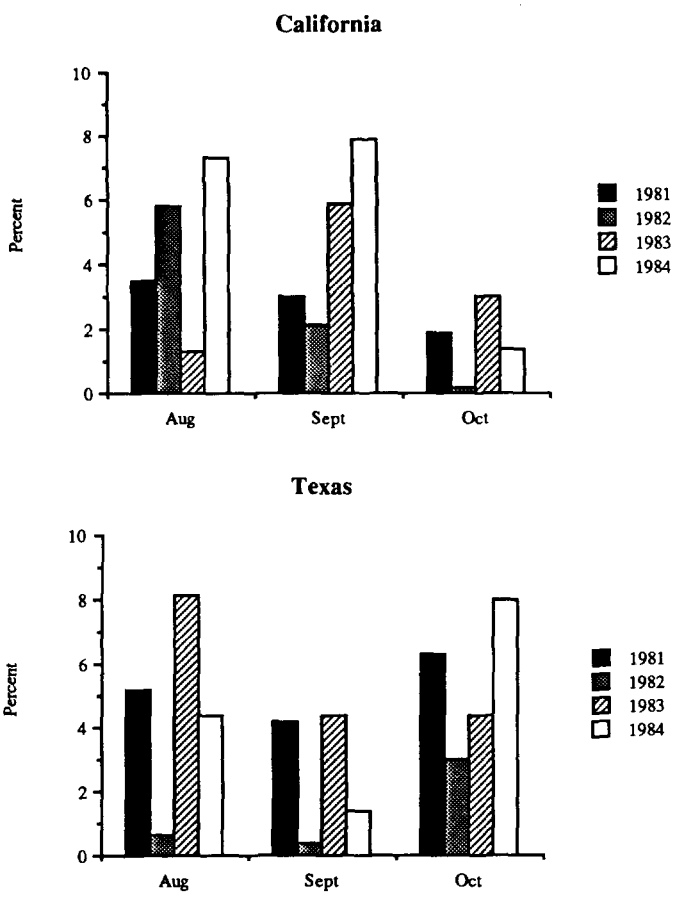

Fig. 5. Comparisons of forecast errors by year within month for California and Texas.

over the years, but new plant condition classes were defined for each year using the procedures outlined previously. The results of this study are given in Table 5 and in Figs 4 and 5. The mean per cent error over the four years was 4·5% in August, 4·7% in September, and 1·6% in October for California, with the comparable figures of 4·6%, 2·6%, and 5·4% for Texas. The results are gratifying, particularly in the light of the large yearly variation in mean annual Texas cotton yields which range from 328 to 465.

RECOMMENDATIONS AND FUTURE RESEARCH

The primary objective of the present study was to investigate the utility of the Markov chain approach in predicting crop yields from large operational data sets. Such 'real world' data sets are characterized by the small number

of time periods at which observations are made, and by the relatively large variation among observations, reflecting the considerable differences among the management and environmental conditions of the fields. Current forecasting procedures in routine use are based on linear regression methodology which requires stringent assumptions, e.g. normality and specified linear equations. These assumptions are widely regarded as unrealistic; however, very few, if any, proven alternative techniques have been available. Indeed, we are not aware of any operational crop forecasting procedure based on similar discrete probability modeling, aside from a previous report on corn forecasting written by one of us (Birkett, 1987). The chief conclusion of this study is that the new method has been successful, as judged from the relatively small forecast errors. In general, statewide forecasts within 5% are deemed exceptional.

The objective yield survey gathers information on plant biometrical characteristics and field management variables. Other variables, describing for example relevant economic, environmental, or remote sensing factors, which might be available for other applications would be easy to include in the analysis outlined.

A number of questions remain concerning possible improvement in the methodology, and they are under current investigation. One question concerns the optimal number of prior years to include in the baseline data set. Usually crop forecasts have been projected from only 3 to 5 years of past data in order to protect against changing technology and economic conditions. However, two characteristics of the present procedure are (1) it is nonparametric and (2) it does not assume *a priori* linear or nonlinear regression equations which may be used for extrapolation. Instead, the procedure may be classified as an adaptive process of matching observed preharvest conditions to historical precedents. Consequently, the procedure may give relatively poor forecasts for individual sampling units with extreme conditions not previously encountered in the baseline data set. The number of sampling units with previously unobserved extreme conditions in a new year under consideration is inversely related to the number of years in the baseline data set. Therefore, in a static technological and economic environment, a larger number of years in the baseline data would tend to improve the forecast by substantially reducing the number of outlier observations.

In the present application, the regression models have a significant year effect which greatly increases the likely number of extreme observations. Nevertheless, the procedure was very successful in forecasting the Texas yields despite their substantial between year variability. In particular, the procedure gave an acceptable forecast of the 1981 mean yield of 465 kg/ha, which exceeded all the other annual means by at least 15%. This success is

due in part to the large within year variability of the Texas sampling units, however, such predictions tend to become less stable when based on just a small number of prior observations. Clearly, it would be very useful to have decision rules determining the optimal size of the baseline data set as a function of three factors; namely (1) within year variability, (2) between year variability and (3) the effect of technological and economic change on the nature of the relationship between variables.

Other questions concern the number of classes which should be used, particularly in relation to R^2 or to the number of observations, n, in each class. We have found that the predictions are quite robust against the number of classes; however, objective criteria could be established to define the number of classes for particular applications.

In summary, we believe that the present methodology has been successful in the present application and would be relatively easy to adapt for many other similar applications. Therefore the Markov chain approach represents a new useful procedure for crop forecasting from operational survey data.

REFERENCES

Birkett, T. (1987). A probability model for Illinois corn yields, *USDA NASS Report No. SRB-87-08*, Washington, DC.

Conover, W. J. (1980). *Practical Nonparametric Statistics* (2nd edn), Wiley, New York.

Matis, J. H., Saito, T., Grant, W. E., Iwig, W. C. & Richie, J. T. (1985). A Markov chain approach to crop yield forecasting. *Agricultural Systems*, **18**, 171–87.

SAS Institute Inc. (1982). *SAS User's Guide: Statistics, 1982*, ed. N. C. Cary, SAS Institute Inc.

US Department of Agriculture (USDA) (1987). Cotton forecasting and estimating models. Section 15 C. In: *1987 Objective Yield Supervising and Editing Manual*. Washington, DC.